NATIONAL AERONAUTICAL ESTABLISHMENT

SCIENTIFIC AND TECHNICAL PUBLICATIONS

AERONAUTICAL REPORTS

Aeronautical Reports (LR): Scientific and technical information pertaining to aeronautics considered important, complete, and a lasting contribution to existing knowledge.

Mechanical Engineering Reports (MS): Scientific and technical information pertaining to investigations outside aeronautics considered important, complete, and a lasting contribution to existing knowledge.

AERONAUTICAL NOTES (AN): Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

LABORATORY TECHNICAL REPORTS (LTR): Information receiving limited distribution because of preliminary data, security classification, proprietary, or other reaons.

Details on the availability of these publications may be obtained from:

> Graphics Section,
> National Research Council Canada,
> National Aeronautical Establishment,
> Bldg. M-16, Room 204,
> Montreal Road,
> Ottawa, Ontario
> K1A 0R6

ÉTABLISSEMENT NATIONAL D'AÉRONAUTIQUE

PUBLICATIONS SCIENTIFIQUES ET TECHNIQUES

RAPPORTS D'AÉRONAUTIQUE

Rapports d'aéronautique (LR): Informations scientifiques et techniques touchant l'aéronautique jugées importantes, complètes et durables en termes de contribution aux connaissances actuelles.

Rapports de génie mécanique (MS): Informations scientifiques et techniques sur la recherche externe à l'aéronautique jugées importantes, complètes et durables en termes de contribution aux connaissances actuelles.

CAHIERS D'AÉRONAUTIQUE (AN): Informations de moindre portée mais importantes en termes d'accroissement des connaissances.

RAPPORTS TECHNIQUES DE LABORATOIRE (LTR): Informations peu disséminées pour des raisons d'usage secret, de droit de propriété ou autres ou parce qu'elles constituent des données préliminaires.

Les publications ci-dessus peuvent être obtenues à l'adresse suivante:

> Section des graphiques,
> Conseil national de recherches Canada,
> Établissement national d'aéronautique,
> Im. M-16, pièce 204,
> Chemin de Montréal,
> Ottawa (Ontario)
> K1A 0R6

DISTANCE MEASURES FOR SPEECH RECOGNITION

LES DISTANCES SPECTRALES POUR LA RECONNAISSANCE DE LA PAROLE

by/par

M. J. Hunt & C. Lefèbvre

National Aeronautical Establishment

OTTAWA
MARCH 1989

DTIC
S ELECTE D
MAY 1 2 1989
E

S.R.M. Sinclair, Head/Chef
Flight Research Laboratory/
Laboratoire de recherches en vol

G.F. Marsters
Director/directeur

89 5 12 013

Summary

This report is concerned with the application of aspects of statistical pattern classification to speech recognition. It presents an extension of linear discriminant analysis to the case where the classes are unknown. This extension provides solutions to the interrelated problems of the design of acoustic representations and spectral distance measures, and allows the efficient combination of heterogeneous sets of parameters. In particular, a representation called IMELDA based on the output of a filter-bank and its changes in time is introduced. Other approaches to distance measures are discussed. It is noted that these other methods lack the ability to make efficient combinations of heterogeneous parameters, and that they require empirical adjustments in order to give good results. Tests indicate that IMELDA provides markedly superior recognition performance compared to the alternatives.

Résumé

Ce rapport traite de l'application des aspects de la classification statistique des motifs à la reconnaissance de la parole. Il présente une extension de la technique de l'analyse discriminante au cas dans lequel les classes ne sont pas connues. Cette extension fournit une solution aux problèmes de la conception des représentations acoustiques et des distances spectrales, les deux problèmes étant inséparables. Elle permet la combinaison efficace de plusieurs séries de paramètres hétérogènes. En particulier, une représentation nommée IMELDA basée sur le rendement d'une banque de filtres et son changement dans le temps est présentée. D'autres approches aux distances sont discutées. On remarque que ces alternatives n'offrent pas la possibilité de faire des combinaisons efficaces de paramètres hétérogènes, et qu'ils ont besoin des ajustements empiriques pour obtenir de bons résultats. Des tests indiquent qu'IMELDA donne une performance en reconnaissance nettement supérieure à celles obtenues avec les approches alternatives.

CONTENTS

1. Introduction

This report describes a method of deriving a linear transformation that can be applied to acoustic parameters for speech recognition in order to improve the speech sound comparison process. It consists of an extension of the pattern classification process known as linear discriminant analysis to the case when the classes are unspecified and may not be discrete.

The method allows several disparate sets of acoustic parameters to be combined into a single, compact representation. Results are reported here with representations that we call IMELDA, generated by combining parameters derived from a mel-scale filter bank.

A close variant of the method described here was first reported at an Acoustical Society of America meeting in 1979 [1]. However, although a full text was produced, only an abstract was published, and that, together with the lack of experimental results at the time, resulted not surprisingly in the work being ignored. The presentation argued that cepstrum coefficients should be given monotonically increasing weights in Euclidean distance calculations, and it provided a method of estimating the values of the weights. Since then, application of weights to cepstral coefficients has become popular. Various theoretical arguments have been advanced for the weighting, but for good experimental results it has usually been necessary to make empirical changes to the theoretical values. Results now show that the method described in principle in 1979 gives exceptionally good results without the need for empirical adjustments. The 1979 text is therefore included as an appendix to this report.

The purpose of this report is to give a tutorial introduction to a method of deriving an acoustic representation with only a brief account of speech recognition tests. A fuller account of recognition tests, covering a wide range of representations is provided elsewhere [2].

2. Pattern Classification and Linear Discriminant Analysis (LDA)

In a typical pattern recognition task, a sample must be assigned to one of a set of known, discrete classes. Values of a set of N parameters are provided for the sample to be classified. The optimum strategy for this problem — namely, Bayes classification — is both well known and obvious: the sample is assigned to the class to which it is most likely to belong. That is, assuming equal *a priori* class probabilities, the sample is assigned to the class whose probability density function is highest at the point in N-space corresponding to the values of the N parameters provided for the sample. This assumes that for each class the probability density as a function of the N parameters is known. In general, however, the probability density functions are not known, and we have to try to estimate them from a finite number of training samples for each class.

If the number of training examples is very large, we can estimate the local class densities directly by k-nearest-neighbor methods, which assume only that the functions are continuous. For somewhat fewer training examples, it is useful to smooth the density estimates over the parameter space using weighted k-nearest-neighbor methods or Parzen functions.

More commonly, however, we either do not have enough training examples to estimate the local densities directly, or else the computational cost of making the estimates for each sample to be classified is considered too high. In these cases, we have to resort to parametric assumptions about the class probability distributions, and generally assume that the distributions are multivariate Gaussian. By computing centroids and covariance matrices for each class, we can estimate the probability densities corresponding to the parameter values of the sample to be classified.

With even fewer training examples, estimates of individual class covariance matrices become unreliable. It is then more effective to assume that classes differ only in their centroids and not in their distributions about those centroids. The covariance information is pooled over all classes to provide a single within-class covariance matrix, W. Note that this matrix is not the same as the covariance matrix, T, representing the covariance of all the training examples without regard to their class, and therefore describing the distribution of the samples about their grand mean. The relation between these two matrices is $T = W + B$, where B is the covariance matrix of the class centroids.

The assumption of identical within-class distributions makes the classification process computationally efficient. A single linear transformation of the measured parameters results in a space in which the log probability of a sample belonging to a given class is directly proportional to the Euclidean distance from the sample to the class centroid. The sample can then be assigned to the class of the closest centroid in the transformed space.

The linear transformation confers spherical symmetry on the within-class probability density functions. The space can therefore be subjected to further rotations without affecting these functions.

In general, the distribution of class centroids will not be uniform in the space designed to make the within-class distributions spherical. By computing a covariance matrix of the centroid coordinates, and carrying out a principal components analysis (i.e. finding the eigenvectors of this matrix) we can obtain an ordered list of orthogonal axes going from the axis along which the dispersion of the centroids is greatest to that along which it is least. If the class centroids are themselves multivariate normally distributed, the ability to discriminate between classes in any direction in the space is given by the variance of the class centroids in that direction divided by the variance of the individual samples about their class centroids. But the latter variance is the same in all directions, so the principal components analysis provides a list of axes going from the one with the greatest discriminating power to the one with the least. The rotation corresponding to the principal components analysis can be combined with the first transform to provide a single linear transformation that can be applied to the original parameters to give a set of transformed parameters. Classification using these parameters is simply a Euclidean distance calculation from the unknown sample to the class centroids, and the parameters are listed in order of their effectiveness. This process is known as *Linear Discriminant Analysis* (LDA).

Because the parameters are ordered by their effectiveness, the computation needed to classify a sample can be reduced by dropping the last few parameters.

Perhaps surprisingly, this process usually improves classification performance. This is probably because the estimates of the class centroids will contain errors. In directions with low classification power, the small differences between the centroids will tend to be dominated by the estimation errors — the "signal-to-noise ratio" will be low.

3. Speech Recognition using Dynamic Programming Template Matching

In speech recognition systems using dynamic programming (DP) template matching, the speech to be recognized and the reference templates with which it is compared are represented as sequences of frames. The frames consist of a set of parameters generally representing the power spectrum sampled every 10 ms or so. By repeating or deleting frames, the sequences being compared are time distorted to find an alignment that maximizes their similarity. Ideally, the *similarity* being maximized should be the *phonetic* similarity of the speech sounds. It is assumed that the alignment will bring together corresponding parts of the words. The alignment chosen is that which minimizes some dissimilarity measure summed over aligned frame pairs across the word. Most commonly, the dissimilarity measure is Euclidean squared distance.

The assumption that seems to underlie this process is that the squared distances correspond to log probabilities. This assumption implies that each frame in a template corresponds to the centroid of the values of the parameters at this location in the word, and that the parameters in the corresponding frames of the individual examples of this word are distributed about the template values according to a multivariate Gaussian distribution with equal variance in all directions. That is, the vector $\bar{x}_k - x_{kj}$ where $\bar{x}_k$ is the vector of template parameter values for the k'th frame and x_{kj} is the corresponding vector for the j'th example of the word, is multivariate normally distributed with spherical symmetry. A consequence of this assumption is that the covariance matrix whose nm'th element is:

$$\sum_j (\bar{x}_{km} - x_{kjn})(\bar{x}_{kn} - x_{kjm})$$

where the subscripts m and n refer to the m'th and n'th parameters, will be diagonal with equal diagonal elements, *i.e.* it will be a multiple of an identity matrix.

If the pairs of frames aligned are assumed to belong to the same class, this matrix can be considered to be a within-class matrix. In making this assumption, we do not have to define what a class is.

In principle, we could compute for each frame in each template a linear transform that would make the Euclidean distance proportional to the log probability that the pair of frames being compared represent equivalent frames in the same word. In practice, there is rarely enough data to obtain accurate distribution estimates for each template frame; and even if there were, the recognition process would be computationally expensive, since each frame-to-frame comparison would require a different transform. Roughly, if a frame consisted of N parameters, the computation needed in the recognition process would be

increased by a factor of N. (Frame-to-frame comparisons account for most of the computation in all but the smallest-vocabulary recognition systems.)

In the previous section we saw that LDA assumes that all classes have the same within-class distributions. We can make the same assumption here and average the within-class covariance estimates over all frames in all templates. The assumption is unreasonable: clearly different speech sounds and different contexts will have different covariance properties; but it is better than the alternative assumptions that will be discussed in the next section. Furthermore, we can also take the second step in LDA and try to estimate a between-class covariance matrix. Exactly what this means when we have not defined the classes is not clear. We can take each frame to be a distinct class with its centroid represented by the template parameter values. Since adjacent frames in steady sounds can be indistinguishable, this would mean that certain "classes" were indistinguishable. Nevertheless, our two matrices will provide a practical measure of the discriminating power of linear combinations of the parameters and will allow us to generate a reduced set of transformed parameters that optimize discrimination with Euclidean distance calculations.

4. Spectral Representations and Distance Measures

A common first stage in a speech recognition system is a bank of around twenty band-pass filters. The filters may be realized directly as analogue or digital filters, or they may be simulated from a Fourier transform. The centre frequencies of the filters are usually approximately equally spaced on a perceptual frequency scale, and their outputs are represented as log energies sampled every 10 or 20 ms.

It is possible to compute frame-to-frame distances directly on the channel log energies, and some systems do so. However, this method is computationally expensive, because it means that all twenty energy values must be stored and compared. Moreover, it can give a poor indication of phonetic similarity because it is sensitive to fine structure. In the lower channels, for example, harmonics of the fundamental may be separated by two channels, particularly for women and children. In this case, phonetically identical sounds spoken on different pitches can be judged to be quite different when the distance is computed channel by channel.

The problem of sensitivity to fine structure can be resolved by using a truncated cosine transform of the sequence of log energies across the channels in each frame. Since a cosine transform is an orthogonal transform, Euclidean distances are unaffected if the twenty channel values are replaced by the twenty transform coefficients. However, the lower-order coefficients are sensitive to the gross structure of the channel energy sequence and the higher-order coefficients are more sensitive to its fine structure. Therefore, by truncating the series at, say, the eighth term we can exclude the fine structure and effectively smooth the spectrum. Moreover, while energy values in adjacent channels are correlated (since spectra are generally smooth) the cosine transform coefficients are largely uncorrelated; that is, the cosine functions are close to a principal components basis set for the spectra (see Figures 1 and 2). This property means that the

lower-order cosine coefficients form an almost optimally efficient representation of the variations between frames. For these two reasons, a cosine transform representation of the filter-bank output (often called a *mel-cepstrum* representation) is widely used.

The most common alternative to a filter-bank front-end in speech recognition systems is linear predictive (LPC) [3] analysis. A cosine transform representation can be easily derived from such an analysis, and Euclidean distance on this representation is now popular with LPC-based systems as well as with filter-bank based systems. In the LPC case, the frequency scale is generally linear, but perceptual frequency scales are also being used [4] at the price of additional computation.

The cosine transform coefficients can be weighted before the frame-to-frame distance is computed. The most trivial weighting, which has been used from the start, is to ignore the coefficient C_0, the coefficient representing total log energy across the channels. Since C_0 is in principle the only coefficient that is affected by changes in gain, setting the weight of this coefficient to zero removes the dependence of the representation on the input level.

More recent and more sophisticated weighting schemes generally apply weights that increase monotonically with coefficient number, at least for the first few coefficients. There are several different motivations for doing this, but the resulting weighting patterns are nevertheless fairly similar.

One motivation is perceptual. Klatt [5] found that the unweighted Euclidean mel-cepstrum distances between sounds correlated well with human judgments of the similarity between sounds as sounds, but correlated poorly with human judgments of phonetic similarity, which is what is needed for speech recognition. He found that a metric that was sensitive to the slope of the smoothed power spectrum gave a much better indication of phonetic similarity. Paliwal [6] pointed out that weighting each mel-cepstrum term by its index (so-called *quefrency weighting*) was equivalent to differentiating the spectrum, and therefore led to a slope-sensitive metric.

A second motivation is based on the properties of speech signals. Giving increasing weight to the higher order coefficients has an effect equivalent to reducing the bandwidths of formants. As a result, spectra recomputed from suitably weighted cepstra show enhanced sensitivity to formants, reduced sensitivity to spectral tilt, and an increased ability to resolve neighboring formants [7].

A third motivation is statistical. It has been claimed [8] that weighting coefficients inversely by their standard deviations should result in a theoretically optimal distance measure. The previous section, however, argued that it is the *within-class* statistics that should be used, not the total statistics. Figure 3 shows the first three principal components of the within-class variance for our database.

When weighting schemes have been tested in recognition experiments [7,8,9], it has generally been found that the best performance is obtained when weighting of the higher components is limited in some empirical way, such as making all weights above a certain coefficient (around the fourth or fifth) equal. Figure 4 shows weights derived from total variances of mel-cepstrum coefficients in a large

multi-speaker database and from within-class variances. The values shown are the reciprocals of the standard deviations normalized to make the weight for C_1 equal to one. (Figure 5 shows for comparison the corresponding weights for the true principal components.) Note that both sets of cepstrum weights rise less quickly than linearly with index number. Moreover, the within-class weights level off around the fifth value, while the weights derived from total variances continue to rise. The empirical expressions may therefore be interpreted as attempts to approximate the statistically derived within-class weights.

5. LDA-derived Transformations with Other Parameters: IMELDA

So far, we have considered only parameters representing the static log power spectrum. But there are other parameters that can be used to represent speech sounds. For example, linear prediction coefficients describe the power spectrum but are not linearly related to it. Parameters can also be derived [10] that represent rates of change in the power spectrum rather than its static properties. Auditory models, such as the one developed at NRC [11], can produce multiple representations that depend nonlinearly on both the static spectrum and on changes in the spectrum. We have seen that the cosine transform is a good approximation to a principal components representation of the log power spectrum, but there is no reason to suppose that it would be suitable for other parameter sets. For representations of linear rates of change in the log power spectrum the cosine transform seems to remain a reasonable approximation to a principal components representation. However, we have no indication as to how to combine the static and dynamic parameters. Most researchers simply append weighted cepstrum coefficients from the dynamic parameters to those from the static parameters. This doubles the number of parameters and thus the storage and computation requirements. Moreover, the relative weights given to the static and dynamic parameters have to be determined empirically with recognition tests.

The approach that we have developed, outlined in Section 2, allows many disparate parameter sets to be combined and a small number of discriminant functions to be derived from them. Using our multi-speaker spoken-digit database, templates are generated by time aligning and averaging together all examples of each digit from all speakers of a given sex. All individual examples from that sex are then re-aligned to the appropriate template to compute a within-class covariance matrix, and a between-class matrix is computed over all the template frames. The approach has been applied successfully to our auditory model [12], and Figure 6 shows the first three discriminant functions derived from a 20-channel mel-scale filter-bank. Note that they are quite different from the functions shown in Figures 2 and 3.

We are currently using our statistical method of generating representations with three parameter sets derived from a twenty-channel mel-scale log power spectrum. The three representations are: the log-power spectrum itself, linear regression coefficients representing the rate of change in log power in each channel, and a notch-like representation generated by taking the log of the sum of the linear energies in pairs of channels spaced two channels apart (e.g. channels 10

and 12). Because our process integrates several **mel**-scale representations using linear discriminant analysis, we have called it *IMELDA*.

Our approach to deriving transformed parameters allows the transformation to be derived for the conditions in which it is to be used. If, for example, it is expected that the test material will be subjected to a particular distortion while the reference material will be undistorted, we can align degraded material to undegraded templates. If the degradation consists of a variation in spectral tilt, application of this degradation during the derivation of the transform will reduce the influence of parameters sensitive to tilt in the discriminant function. In our experiments, we have been interested in both tilt and added noise, so we have computed a transform using a combined within-class covariance matrix resulting from alignment of tilted, noisy and undegraded speech. Such a "multi-condition" transform gives better performance over the three test conditions than a transform derived from a single condition. Moreover, its performance on undegraded test speech is only slightly worse than that of a transform derived purely from undegraded speech.

Without degradations, the mean of all the frames aligned to a template frame should correspond to the template frame, because the template frame was derived in this way. After degradation, however, the means will no longer be equal. This invalidates one of the assumptions on which our transformation is based, namely that the individual frame parameter values will have a multivariate Gaussian distribution about the corresponding template frame parameter values. It is possible to correct for the shift in the mean in the computation of the covariance matrix, but since we do not shift the templates in the recognition experiments, it is not obvious that we should make this correction: a parameter whose mean shifts is less useful and should be given less weight. We have compared recognition performance when the correction for mean shift was applied and when it was not, and we find that it is indeed better not to apply the correction.

For the auditory model, there is no conventional alternative to our method of deriving a transform using LDA, though we have compared it with using principal components derived from the template frames and found the latter to be inferior. For filter-bank representations, however, we can make direct comparisons between IMELDA and alternatives used elsewhere. Table 1 shows a comparison in an isolated-word digit recognition test between IMELDA, mel-cepstrum, and a representation developed at NTT using mel-cepstrum and weighted regression coefficients for the change of cepstrum coefficients over time ("δ-cep"). Dynamic programming template matching was used with a single averaged template per digit and a 19.2 ms frame rate. The word boundaries were marked manually.

As we noted earlier, the NTT representation has two parameters to be set empirically. The value of one of these parameters could be taken directly from an NTT publication, while we set the other to optimize performance in speaker-independent tests. This procedure implies a slight bias in favor of the NTT method. Nevertheless, the table shows that the NTT representation was far less effective than IMELDA. The magnitude of improvement of the NTT

representation over a conventional mel-cepstrum representation is comparable with that reported by the NTT researchers, which leads us to conclude that we have accurately implemented their method. Alternative weighting schemes proposed elsewhere have shown performance improvements over unweighted cepstrum that are comparable in magnitude with that found for the NTT representation, and therefore much smaller than IMELDA.

It might be objected that IMELDA uses an additional set of input parameters in the notch-like representation, and that this makes the comparison with the NTT representation unfair. Table 1 therefore also shows results with a reduced IMELDA in which the notch was excluded, and an even simpler representation in which only the twenty log channel energies ("LCE") were used as input parameters to the LDA. Even this simplest representation performs enormously better than the representations that do not use LDA.

The mel-cepstrum results were obtained using eight parameters (C_1 to C_8). There seems to be little advantage in going to a larger number of coefficients. With the NTT representation, however, we found that results were improved by using twelve mel-cepstrum and twelve δ-cepstrum terms, so these are the results shown. For comparison, the simplest LDA results and the reduced IMELDA results are shown for both eight and twelve parameters.

Table 2 shows the results of similar tests with a connected-word recognition system. Tests have been carried out only for mel-cepstrum and IMELDA.

6. Use of IMELDA with Other Speech Recognition Techniques

The IMELDA representation is derived using DP time alignment, and it has been tested in systems using DP template matching. Hidden Markov Modeling (HMM) is a popular alternative, and indeed the most effective current speech recognition systems are generally HMM systems. However, given the large improvement in DP template matching performance with IMELDA, it is not obvious that current HMM systems would still have better performance. HMM systems frequently use a cepstrum representation, sometimes augmented with a δ-cep representation. It is therefore natural to ask whether an IMELDA representation could be used. Theoretically, HMM should, through the B matrix, be capable of deriving an ideal distance measure for each hidden state. In practice, however, there is never enough data to estimate the parameters adequately. HMM systems are forced to make simplifying assumptions: for example, that the covariance matrices are diagonal, or that all variances are equal. The assumptions underlying IMELDA may lead directly to better HMM performance, or may at least provide a better starting point for HMM estimations of state-specific variances.

Neural networks are widely felt to be serious competitors to DP template matching and HMM for future speech recognition systems. Here, the applicability of IMELDA is less evident. Small-scale experiments [13] have not found any difference in performance when input parameters were subjected to linear transformations such as a cosine transform or LDA. These results must be regarded as preliminary, however, particularly since the multi-layer-perceptron system used was found to give much worse recognition performance than our

simple DP template matching system on the same spoken-digit database.

7. Conclusions

This report has shown how linear discriminant analysis can be extended to the case when the classes are unknown and possibly non-discrete. This extension has led to a technique for computing distance measures for speech recognition that is theoretically based and needs no empirical adjustment. It nevertheless appears to outperform alternative distance measures in dynamic programming template matching systems even when empirical adjustments are applied to them.

The technique has the unusual property of being able to combine diverse sets of parameters. In particular, the representation known as IMELDA has been found to be particularly effective for robust speech recognition.

The transformations derived by the technique described in this report are likely to be effective in HMM-based speech recognition systems as well as those using dynamic programming template matching, but possibly not in systems using neural networks.

Acknowledgments

References

1. M.J. Hunt, "A Statistical Approach to Metrics for Word and Syllable Recognition," Fall 79 Meeting, Acoust. Soc. America, Salt Lake City, abstract published in *J. Acoust. Soc. America,* Vol 66, pp.S535-536, 1979.

2. M.J. Hunt & C. Lefèbvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech," *Proc. IEEE Int Conf. Acoustics, Speech & Sig. Proc., ICASSP-89,* Glasgow, Scotland, May 1989.

3. J.D Markel & A.H. Gray, *Linear Prediction of Speech,* Springer-Verlag, Berlin, 1976.

4. H. Hermansky, "An Efficient Automatic Speaker-Independent Speech Recognition by Simulation of some Properties of Human Auditory Perception," *Proc. IEEE Int Conf. Acoustics, Speech & Sig. Proc., ICASSP-87,* Dallas, April 1987, pp. 1159-1162.

5. D.H. Klatt, "Prediction of Perceived Distance from Critical Band Spectra. A First Step," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-82,* Paris, May 1982, pp. 1278-1281.

6. K. Paliwal, "On the Performance of the Quefrency-Weighted Cepstral Coefficients in Vowel Recognition," *Speech Communication,* 1982, Vol. 1, pp. 151-154.

7. F. Itakura & T. Umezaki, "Distance Measure for Speech Recognition based on the Smoothed Group Delay spectrum," *Proc. IEEE Int Conf. Acoustics, Speech & Sig. Proc., ICASSP-87,* Dallas, April 1987, pp. 1257-1260.

8. Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *Proc. IEEE Int Conf. Acoustics, Speech & Sig. Proc., ICASSP-86,* Tokyo, April 1986, pp.761-764.

9. B.H. Juang, L.R. Rabiner & J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *Proc. IEEE Int Conf. Acoustics, Speech & Sig. Proc., ICASSP-86,* Tokyo, April 1986, pp.765-768.

10. K. Aikawa & S. Furui, "Spectral Movement Function and its Application to Speech Recognition," *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-88,* New York NY, April 1988, Vol 1, pp. 223-226.

11. M.J. Hunt and C. Lefèbvre, "Speech Recognition with the NRC Auditory Model," *Proc. Canadian Conference on Electrical and Computer Enₑ neering,* Vancouver, Nov. 1988, pp. 135-138.

12. M.J. Hunt & C. Lefèbvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model," *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-88,* New York NY, April 1988, Vol 1, pp. 215-218.

13. W.C. Treurniet, M.J. Hunt, C. Lefèbvre & Z. Jacobson, "Phoneme recognition with a neural network: comparisons of acoustic representations including those produced by an auditory model," *Intl. Neural Network Soc., First Annual Meeting,* Boston, Mass, September 1988.

Test material	Representation	(# of coefs.)	Errors (%) SD	SI
Undegraded	mel-cepstrum	(8)	1.1	5.8
	mel-cepstrum + δ-cep	(24)	0.6	3.6
	LCE & LDA	(8)	0.4	1.8
	LCE & LDA	(12)	0.2	1.5
	reduced IMELDA	(8)	0.3	1.1
	reduced IMELDA	(12)	0.2	0.8
	IMELDA	(12)	0.1	0.3
White noise SNR = 15 dB	mel-cepstrum	(8)	23.9	36.6
	mel-cepstrum + δ-cep	(24)	9.7	19.9
	LCE & LDA	(8)	1.1	5.3
	LCE & LDA	(12)	1.2	5.6
	reduced IMELDA	(8)	1.3	3.3
	reduced IMELDA	(12)	0.9	3.0
	IMELDA	(12)	0.6	1.9
Tilted 6 dB/oct.	mel-cepstrum	(8)	75.4	77.3
	mel-cepstrum + δ-cep	(24)	62.5	71.9
	LCE & LDA	(8)	0.5	4.5
	LCE & LDA	(12)	0.4	3.7
	reduced IMELDA	(8)	0.4	1.9
	reduced IMELDA	(12)	0.4	1.7
	IMELDA	(12)	0.0	0.3

Table 1. Speaker-dependent (SD) and independent (SI) quasi-isolated-word recognition results on 1346 digits from nine male speakers. The numbers in parentheses represent the number of coefficients used in the Euclidean distance calculation.

	test speech	repres- entation	subst.	ins.	del.	tot.	rate %
speaker	undegr.	. IMELDA	0	0	3	3	0.2
		mel-cep.	3	26	23	52	3.8
dep.	noisy	IMELDA	4	0	4	8	0.6
		mel-cep.	316	9	59	374	27.7
	tilted	IMELDA	0	0	1	1	0.1
		mel-cep.	509	74	521	1104	81.7
speaker	undegr.	IMELDA	7	0	17	24	1.8
		mel-cep.	70	42	23	135	10.0
indep.	noisy	IMELDA	33	13	10	56	4.1
		mel-cep.	445	15	65	525	38.8
	tilted	IMELDA·	7	0	12	19	1.4
		mel-cep.	480	98	593	1171	86.6

Table 2. Speaker-dependent and independent connected-word recognition results on 1352 digits from nine male speakers.
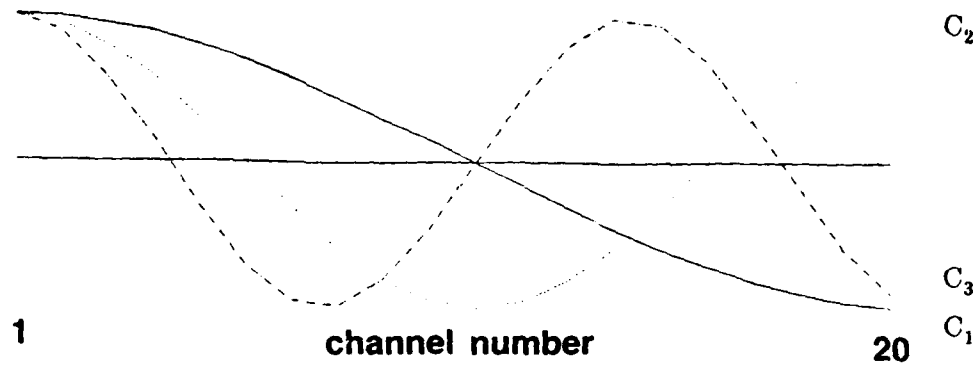
Figure 1. The cosine basis functions corresponding to C_1 (continuous line), C_2 (dotted), and C_3 (dot-dash). The basis function corresponding to C_0 is simply a constant.
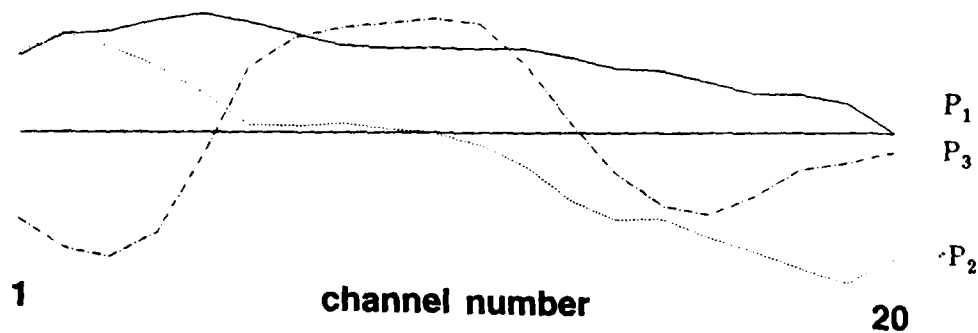

Figure 2. The first three principal components of the total variances in the log energies in our twenty-channel mel-scale filter-bank for our database.
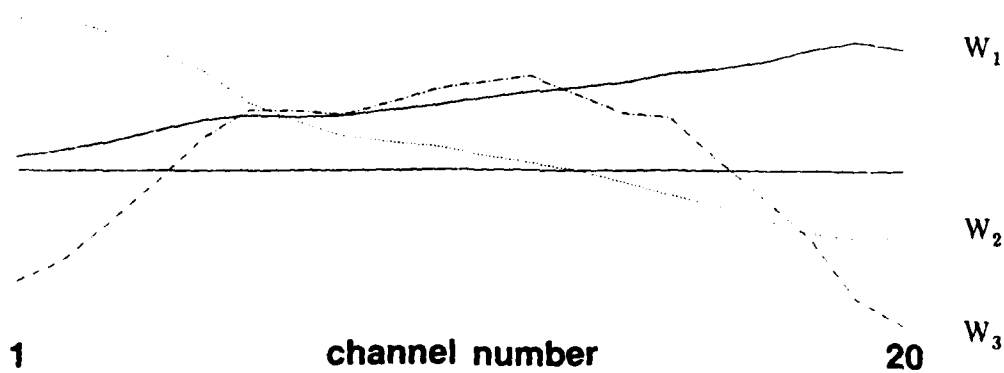

Figure 3. The first three principal components of the *within-class* variances in the log energies in our twenty-channel mel-scale filter-bank for our database.
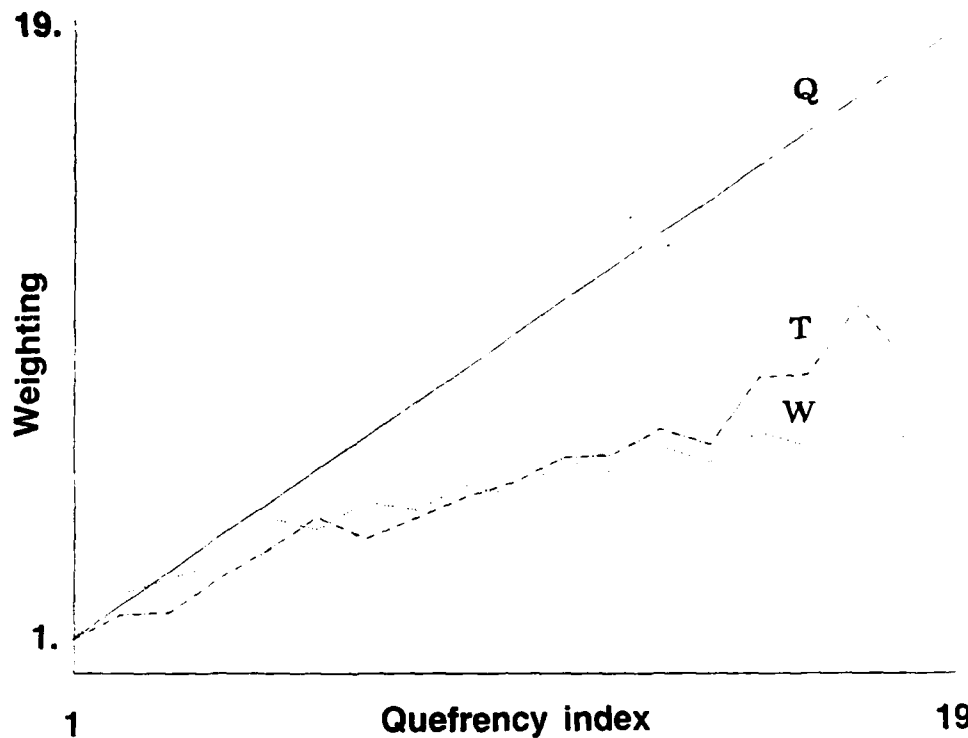
Figure 4. Mel-scale cepstrum coefficient weights computed from the square roots of the total variances (dot-dash and "T"), and of the corresponding values for the within-class variances (dotted and "W"), and from simple quefrency weighting (continuous line and "Q"). Note the slower increase above C_5 in weights derived from the within-class data.
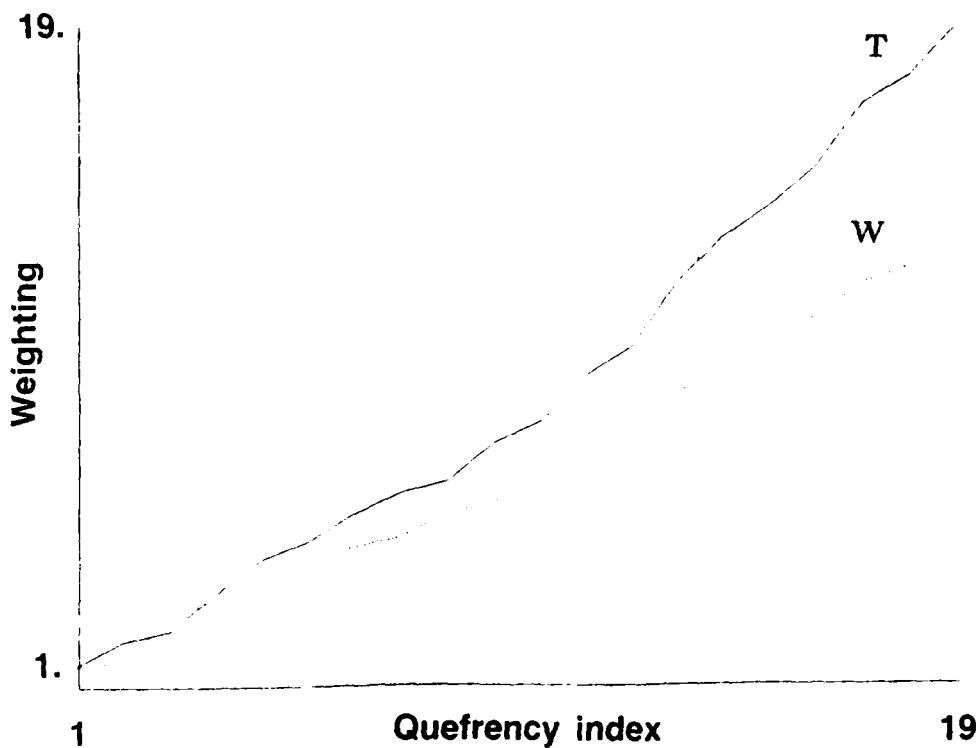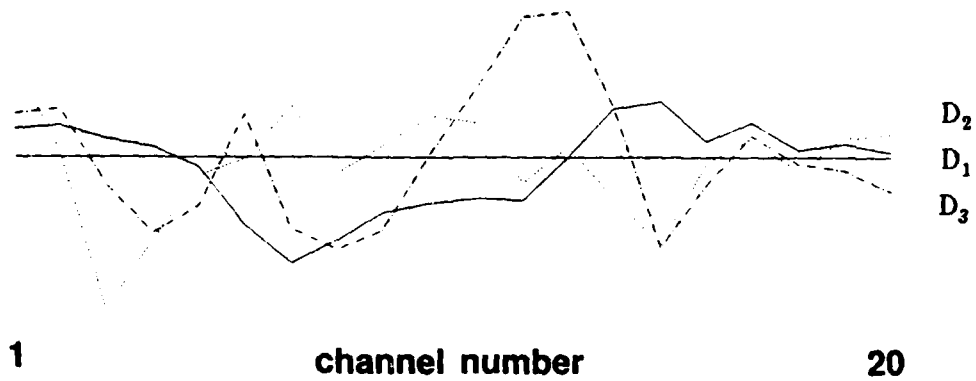

Figure 5. Weights computed as in Figure 4 but for principal components of the total (continuous line and "T") and within-class (dotted and "W") variance data.

Figure 6. The first three discriminant functions derived from the log-channel energies for our twenty-channel filter-bank for our database. Note that the functions are much less smooth than those shown in Figures 1-3: smooth features account for much of the variance, but relatively less of the discriminating information in log power spectra.

APPENDIX

A STATISTICAL APPROACH TO METRICS FOR WORD AND SYLLABLE RECOGNITION

Melvyn J. Hunt (Bell-Northern Research, 3 Place du Commerce,
Verdun, Quebec, Canada, H3E 1H6)

ABSTRACT

Time-warping pattern-comparison algorithms are widely used in speech recognition. Two words or syllables being compared are described by a series of frames each containing values of a set of acoustic parameters. After time alignment, the squared distance between the patterns is summed over the parameters within a frame and then across frames. The sum obtained is assumed to be proportional to the log probability of the two patterns having the same identity. This assumption is generally invalid, but it may be made substantially true by analyzing the variability between different examples of the same syllable and adjusting the metric accordingly. Variability is estimated both as a function of frame position within the syllable and as a function of the acoustic parameters. In the latter case, within- and between-class covariance matrices can be estimated and standard linear discriminant analysis methods applied. This permits the combination of disparate acoustic parameters into a single distance measure. The possibility is considered that different speed sound classes require different metrics, and the problem of interframe correlation is discussed.

1. INTRODUCTION

This paper is concerned with the distance measures used in the comparison of unknown words or syllables with stored reference forms in the dynamic programming pattern matching algorithm. The work forms part of an effort to achieve speaker-dependent continuous speech recognition by first dividing the speech into syllables and then matching the syllables consecutively against stored templates [Hunt, Lennig and Mermelstein 1980].

The templates we use are composites formed by warping together several utterances of the same syllable. Our symmetric, unconstrained formulation of the matching algorithm is particularly well suited to the production of such composites.

Our speech material is digitized with an 8 KHz sampling rate and represented by the first seven mel-scale cepstrum coefficients computed every 6.4 ms. The general approach described here is not, however, confined to this form of representation of the speech signal nor to speaker-independent systems, and the central concern of the paper is with the general approach and not with its application to our particular system. The experimental results that are quoted were obtained from sets of sentences specifying dates and times recorded in English by a male speaker and in French by a female speaker.

The syllable matching algorithm returns a squared distance between an unknown syllable and a composite reference template. For the purpose of recognizing individual syllables the only requirement on the distance is

that it should be monotonically related to the probability that the unknown belongs to the same syllable type as the template. However, when a sequence of syllables is to be recognized, the information coming from the individual syllables has to be combined. If the information regarding individual syllables is considered to be independent, then the appropriate combination is to multiply together the individual probabilities or, equivalently, to add the log probabilities. It is therefore desirable that the distances output by the syllable comparator should be proportional to log probabilities.

The statistical distribution for which log probability is proportional to squared Euclidean distance from the mean is an uncorrelated multivariate normal distribution in which the variances of all variables are equal. In the syllable comparator, the composite template can be regarded as an estimate of the 'mean', and the individual utterances as samples distributed about the mean. We would like to transform the distance measures to give the required log probability property. Note that it is 'within-class' distributions that matter and not the total distribution over all syllable types.

The standard method of carrying out dynamic programming template matching is to find the path linking the two ends of the two words or syllables being matched which has the smallest sum of local squared distances along the length of the path. The local squared distances are obtained by summing the squares of the differences of the corresponding parameters in the pair of frames linked at each point in the path. These parameters describe the spectrum, and may, for example, be channel amplitudes or, as in our case, cepstrum coefficients (we exclude from this discussion the Itakura metric [Itakura 1975]).

Consider the case of a composite template being matched against one of the utterances from which it was constructed. Provided the utterances have been checked to insure that they consist of an identical sequence of phonemes, and provided, as in our case, that there are no constraints on the form of the time warping, then all aligned pairs of frames between the template and individual utterance should represent the same speech sound. Hence, any differences between parameter values in pairs of frames must be regarded as within-class differences resulting from the intrinsic variability of phonemically identical speech sounds. For our assumption to be true that the summed squared differences along the best path measures a log probability, certain assumptions have to hold:

i)   that the within-class scatter is described by a multivariate normal distribution.

ii)   that differences between nearby frame pairs are uncorrelated.

iii) that the variation is identical for all speech sounds in all positions in all syllables.

iv) that the variation is identical and uncorrelated for all parameters within a frame pair.

We take the first assumption to be approximately true. The second assumption is patently untrue: differences in adjacent frame pairs of sustained speech sounds will be highly correlated. We cannot so far claim any success in dealing with this problem.

The third and fourth assumptions seem unlikely to be true. They can be checked in principle by computing variances and covariances for every parameter of every frame in each syllable in the inventory. For a small inventory with many examples of each syllable, such computation would be possible. Alternatively, systems which classify frames into phoneme-like segments could lump together information from equivalent segments. Neither of these options is open to us: we have a large inventory of syllables with relatively few repetitions of each syllable, so statistically reliable estimates for individual frames cannot be made; and we do not attempt phoneme classification - in any case, a dubious and error-prone procedure.

2. WITHIN-CLASS VARIANCE AS A FUNCTION OF CEPSTRUM COEFFICIENT

In estimating within-class variation we are forced into making some generalizing assumptions. Such generalizations have respectable precedents: in automatic speaker recognition, for example, there is rarely enough data to estimate variability separately for each speaker, so speakers are assumed to differ only in the values of the means of the parameters being used to characterize them. Variation about the means is assumed to have an identical multivariate normal distribution for every speaker, and the within-speaker variation information is pooled across all speakers.

In a similar spirit, we have assumed that information about within-class variance and covariance can be pooled across all frames of all syllables. We then obtain a within-class covariance matrix whose elements, c_{ij}, are given by,

$$c_{ij} = \frac{1}{N} \sum_{k,m,n} (x_{ijmn} - \overline{x}_{ikm})(x_{jkmn} - \overline{x}_{jkm})$$

Where x_{ikmn} is the value of the i'th cepstrum coefficient number in the k'th frame of the n'th example of the m'th syllable type, and $\overline{x}_{ikm}$ is the corresponding value of the aligned frame of the template of the m'th syllable type. The factor N has to take into account the loss in degrees of freedom from the fact that the template values had been estimated from the data. Consequently, in making a count for N, each time a term was added to the sum for c_{ij}, N was incremented not by 1 but by $(t-1)/t$ where t is the number of tokens used to create the template. Note that if x_{ikm} in the expression c_{ij} were replaced by $\overline{x}_i$, the overall mean value of the i'th cepstrum coefficient, c_{ij} would be an element of the total covariance matrix.

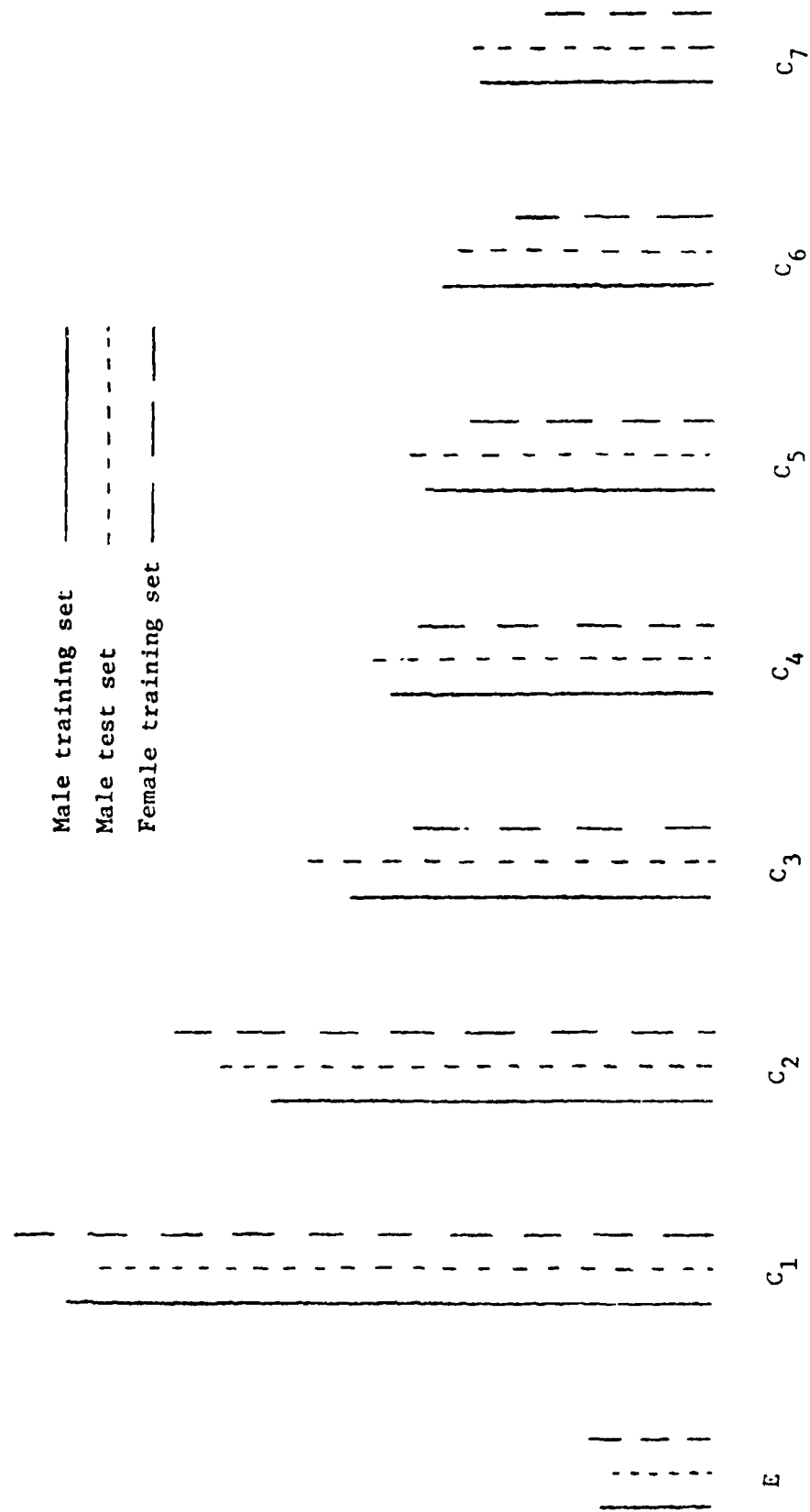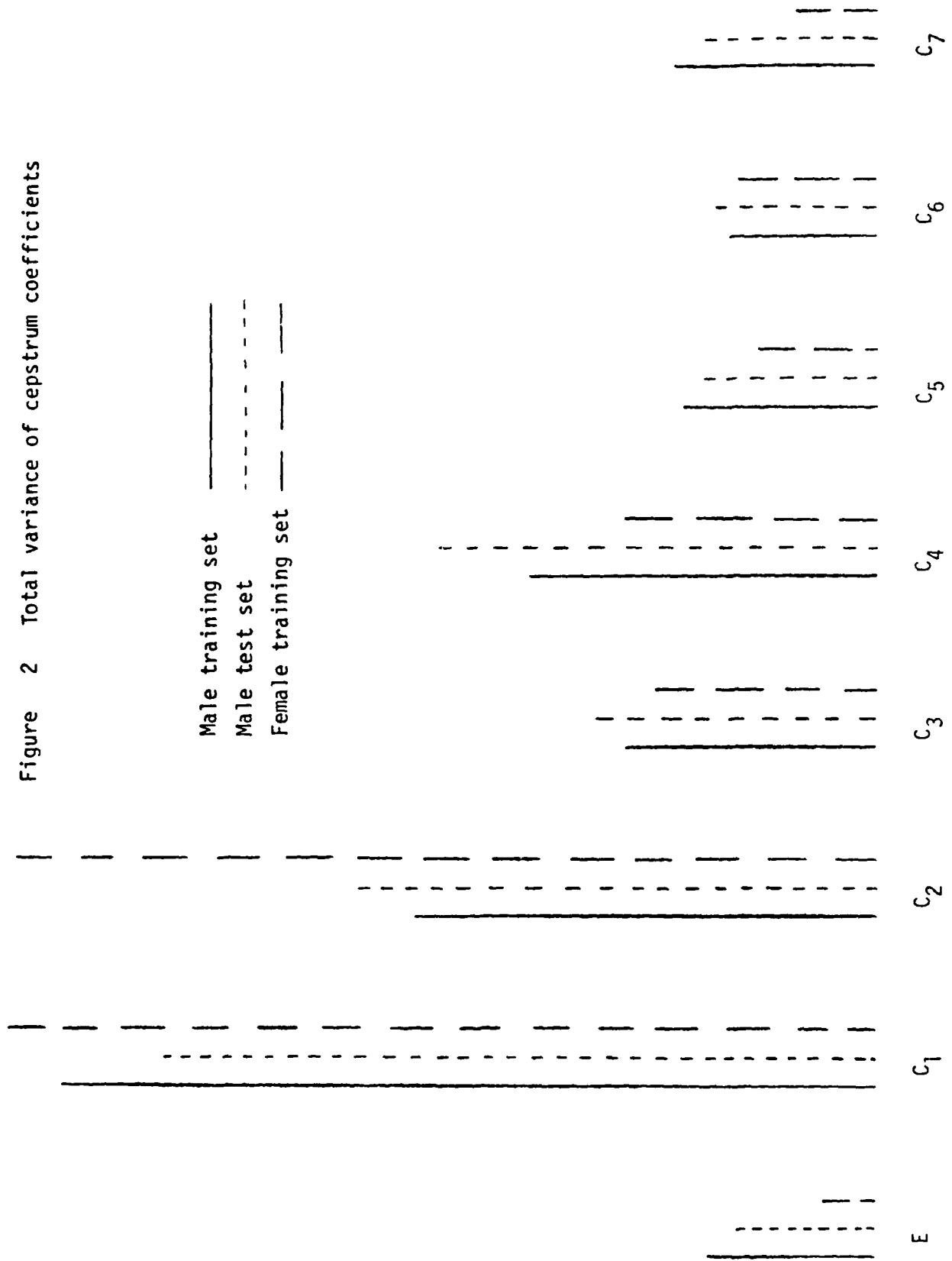Figure 1 Within-class variance of cepstrum coefficients

Figure 2 Total variance of cepstrum coefficients

It is well known that the total covariance matrix is diagonal, i.e. that there is no overall correlation between the cepstrum coefficients. Our own experiments confirmed this result. It does not follow that the within-class covariance matrix will also necessarily be diagonal, but in fact that turns out to be the case. This is convenient because it means that the optimum representation can be achieved by simply dividing the cepstrum coefficients by the square roots of their within-class variances. If we had started off with channel amplitudes there would have been within-class correlations, but we could have ended up with the same optimum representation by applying standard linear transformation techniques to the original parameters in a manner defined by the elements of the (nondiagonal) within-class covariance matrix.

As can be seen in Figure 1, the within-class variances decrease monotomically with increasing cepstrum number. The variances are remarkably consistent across the two speakers and across the two recording sessions for the male speaker. Figure 2 shows the correspoonding values for the total variance. It can be seen that these values generally decrease sharply with increasing cepstrum number, but that the decline is not monotonic. The values are much less consistent across the three data sets than are the within-class variances. Previous attempts at scaling [Pols 1977] have used measures of total variance.

Since it is the relative values of variance which matter for scaling purposes, the values of total and within-class variance shown in the figures has been normalized by their sum across all the coefficients.

The bars marked E in the figures represent log loudness. Although the cepstrum coefficients themselves are not correlated with each other, loudness is correlated strongly with the cepstrum coefficients, particularly C_1. The main reason for this seems to be that voiced sounds are generally much louder than voiceless ones.

The first estimate of the variances has to be made using unscaled cepstrum coefficients. This affects both the construction of the templates and the alignment used in the variance estimation. When this variance information is used to scale the cepstrum coefficients, the resulting alignments in the template generation and in the re-estimation of the variances will be somewhat different. We found that the re-estimated values of the variances did not differ from the initial estimates by more than 10%, and the changes in subsequent iterations of this process were negligible.

3. WITHIN-CLASS VARIANCE AS A FUNCTION OF FRAME POSITION IN THE SYLLABLE

So far, then, we have worked with the assumption that the variance properties of the cepstrum coefficients can be usefully investigated

independently of the other factors which might affect frame variability. In a similar way, we have also made the assumption that within-class variability as a function of position within the syllable can be usefully investigated independently of other considerations. Composite templates are again matched against the examples which went to make them up. The average distance is computed between aligned frames as a function of the number of frames from each end of the syllable. This is carried out up to ten frames in from each end of the syllable, except that for syllables shorter than twenty frames the process stops when the center of the syllable is reached. The count of number of frames from the end could be made on the template or on the token being matched against it, but in keeping with the symmetric formulation of our matching algorithm we define the number of frames from the end of the syllable as the average of the number for the template and that for the token (our formulation of ensures that this average is always an integer). This symmetric definition has the useful property that a weighting applied in this way does not affect the alignment chosen, since at each stage the alternative steps which the algorithm chooses between will all receive equivalent weighting.

Figure 3 shows variability as a function of syllable position for the three sets investigated. Since, again, it is relative variability that matters, the curves have been normalized to be close to each other near the center of the syllables. It can be seen that the variability is up to three times greater at the syllable edges than in the center. There are probably several factors which contribute to this effect. The most obvious is that minor errors in the placement of syllable boundaries will cause a few frames at a syllable edge to be deleted or a few frames from an adjacent syllable to be added. The second possible factor is coarticulation around syllable edges from the adjacent syllable. A third factor is that the alignment path is much more constrained near syllable edges (for example, the first template frame must be aligned to the first token frame). Finally, the realization of the kinds of speech sounds occurring most often at syllable edges may be less consistent than those that are characteristic of syllable nuclei. This possibility is discussed further below.

The scaling by position is applied during recognition by dividing all local squared distances by the relative variability value found for the training set. Local distances more than ten frames inside the syllable are left unscaled.

The scaling by position and the scaling of the cepstrum coefficients would be expected to interact with each other. We do find some interaction, but it is quite small. In particular, the frame-number/ variability curves are very insensitive to scaling of the cepstrum coefficients.

Figure 3 Within-class variance as a function of frame position within the syllable
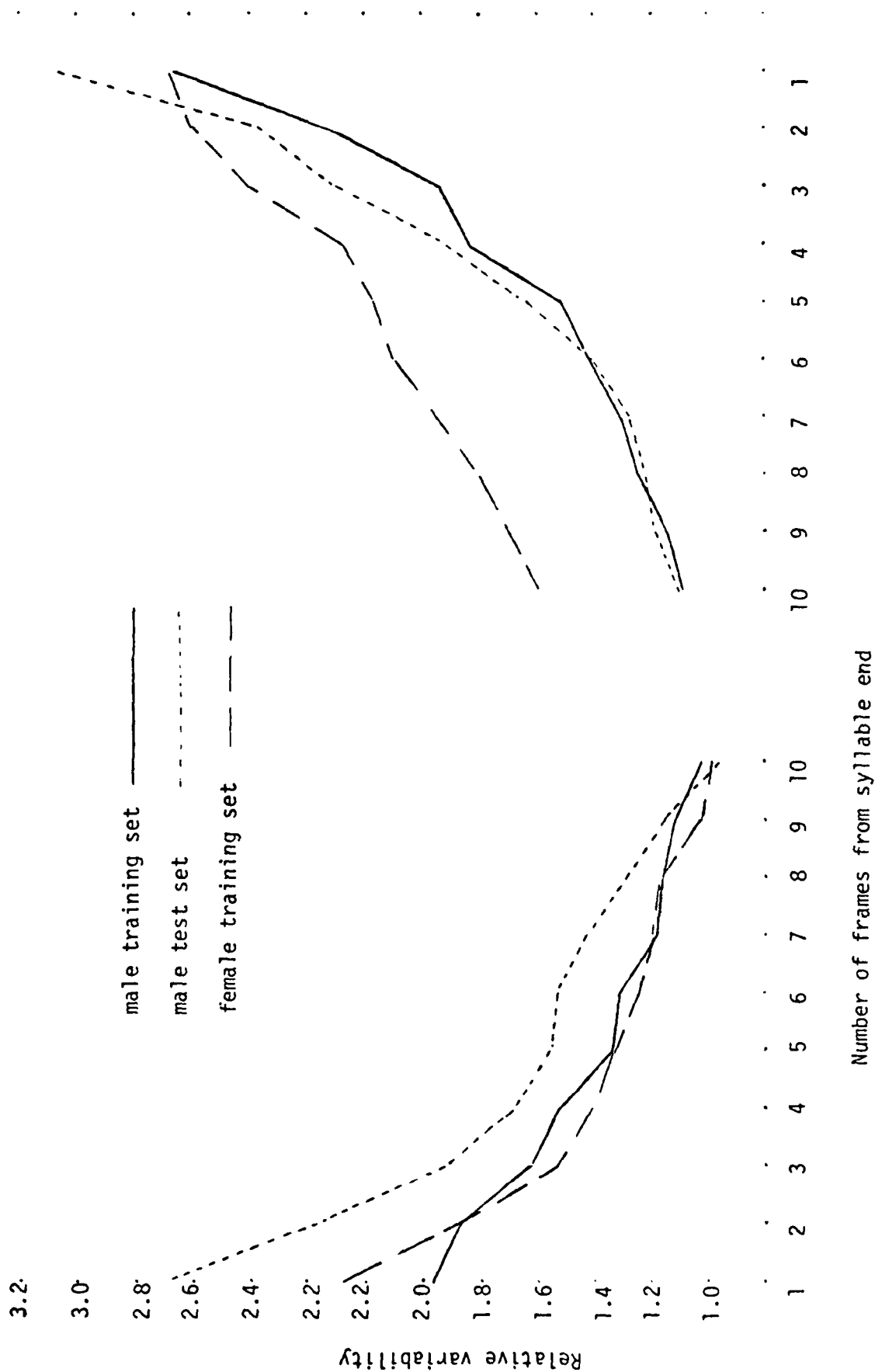
Implementation of these two forms of scaling approximately halved the recognition error: out of the 59 sentences in the English test set the number of misrecognized sentences fell from seven to three, and when the recognition strategy was modified to increase the error rate by retaining only the single most promising sentence hypothesis, there were 21 misrecognized sentences without scaling and 15 with scaling.

4. WITHIN-CLASS VARIANCE AS A FUNCTION OF SPEECH SOUND CLASS

The third factor we have considered which might affect within-class variance is the class of speech sound. We were resolutely opposed to the rigid classification of frames into speech sound classes; rather, we envisaged having a continuously variable metric whose form was a function of the cepstrum coefficients of the frames being compared. It seemed, however, that the easiest way to investigate the idea was to compute two covariance matrices with contributions from each frame comparison being added to one matrix or the other depending on which side of a boundary the value of one of the cepstrum coefficients lay. For example, we tried computing two covariance matrces one of which received contributions when the average values of C_1 for the pair of frames being compared aly about the overall mean for C_1 while the other received contributions when the average value lay below the mean. Similar experiments were carried out using the mean values of other cepstrum coefficients and linear combinations of them. In no case was there any marked difference between the pairs of matrices.

We then plotted histograms of the values of the cepstrum coefficients occurring in the English training material. All of the distributions were unimodal except for C_1, which was clearly bimodal (see Figure 4). The dip between the two peaks occurs around the point that we used in the syllabifier to distinguish voiced from voiceless segments, and it is clear that the two peaks correspond to voiced and voiceless sounds. A boundary placed between the two peaks resulted in a pair of covariance matrices which were markedly different. The matrix resulting from voiceless sounds is a scaler multiple of that for voiced sounds with a scaling factor of around 2.5. The results for the English test material were essentially identical.

It looked, then, as though voiceless sounds had greater within-class variability than voiced ones. The scalar relationship between the two diagonal matrices meant that the differences in variability as a function of C_1 could be represented by the values of the traces of the matrices i.e. by the average squared Euclidean distance between a pair of aligned frames as a function of the mean value of C_1 for the pair. What we expected to see was the variability uniformly increasing as C_1 moved into the voiceless region. This did not happen. As Figure 5 shows, the variability peaks at the voiced/voiceless boundary. The most probable explanation seems to be the following. The largest and

Figure 4   Histograms of values of first cepstrum coefficient

male training set —⊙—

male test set --×---×--

Value of $C_1 \times 10^{-3}$ (arbitrary units, but consistent with Fig 6)
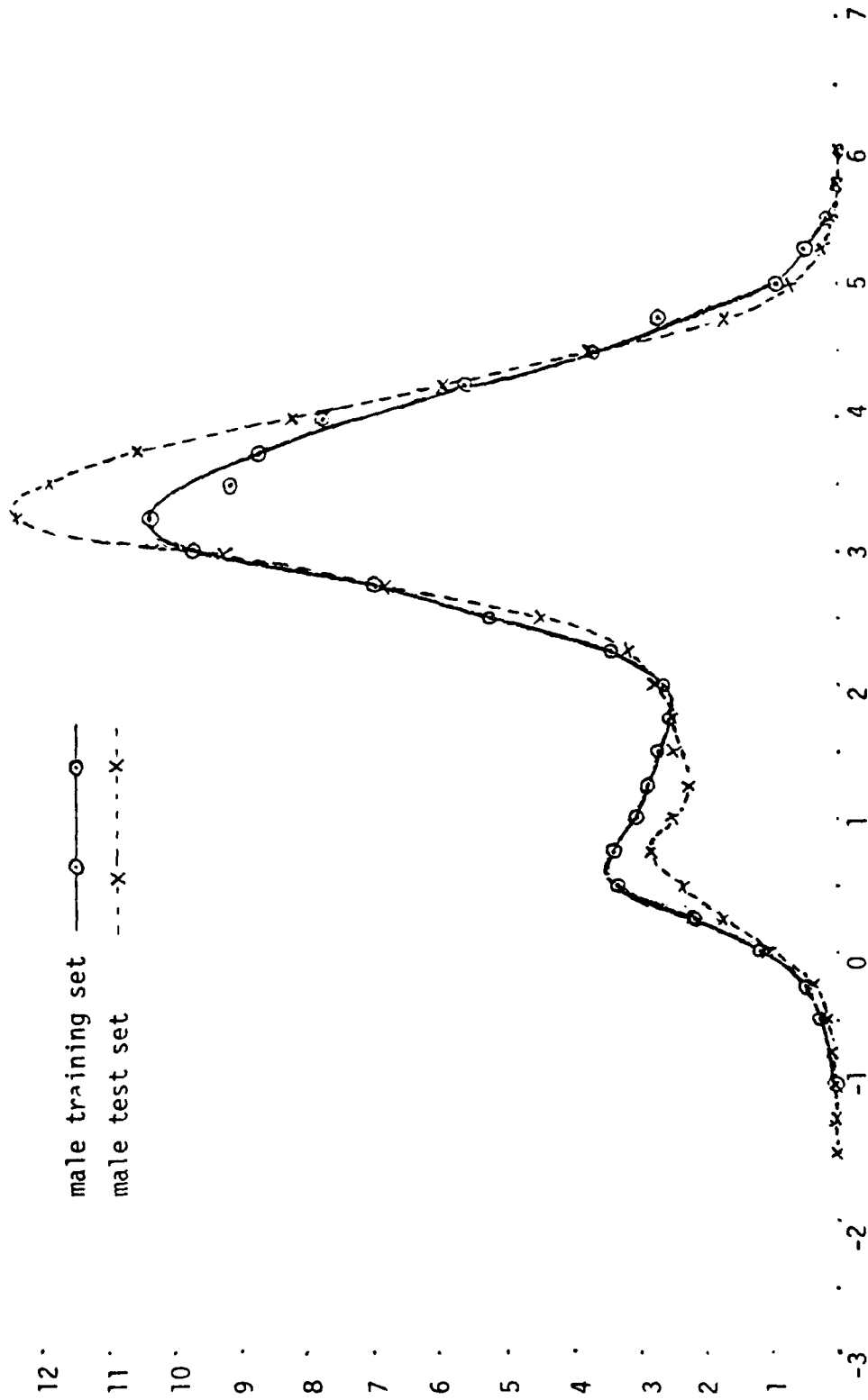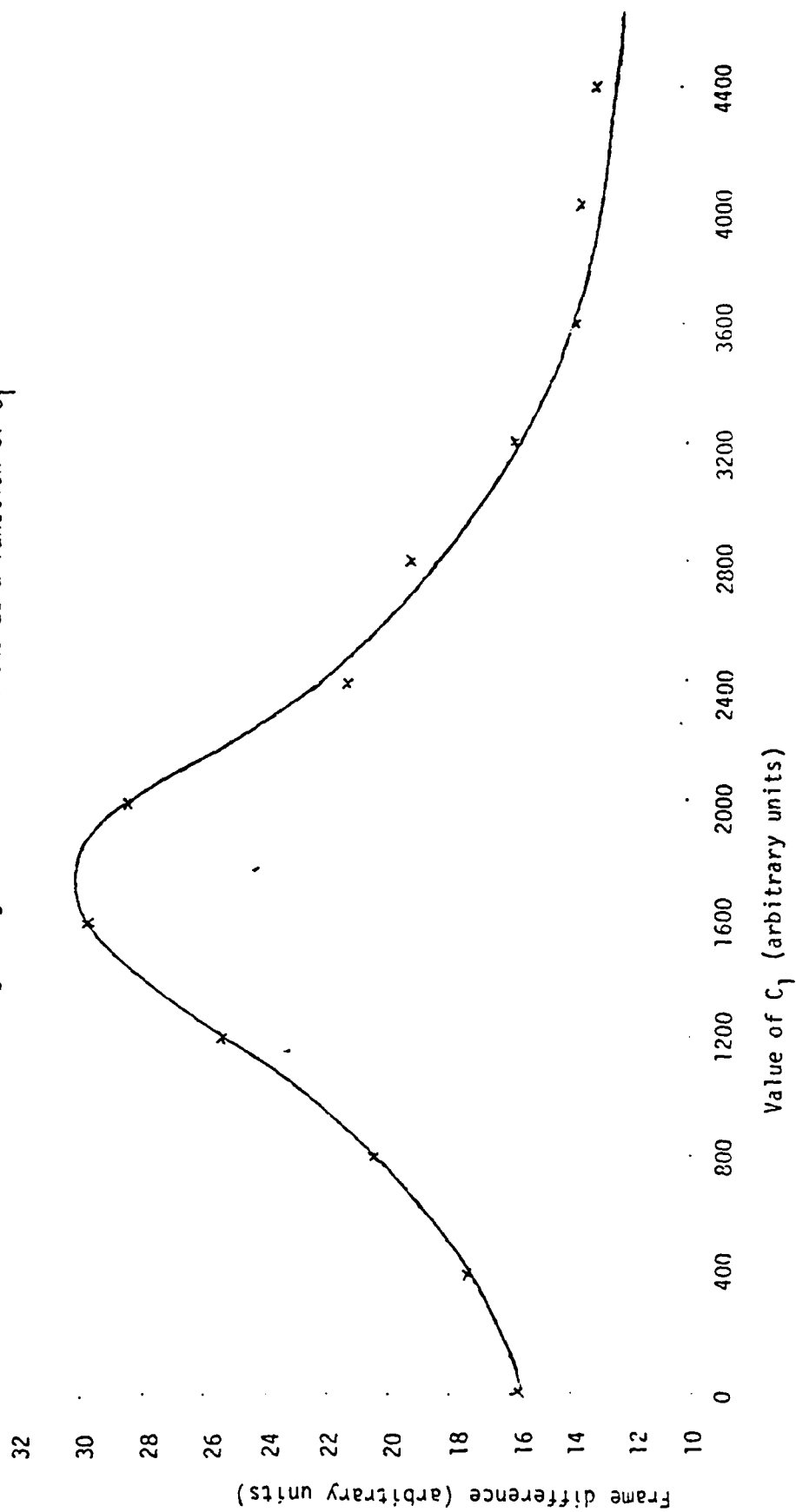
Fig. 5   Average aligned frame differences as a function of C_1

most sudden spectral change that can occur is at the onset and offset of voicing. However well two syllable productions are time-aligned, the quantization into discrete frames means that at a voiced/voiceless transition point a voiceless frame in one production will often have to be matched to a voiced frame in the other production. The mean value of voiced and voiceless values, and the spectral distance between the two frames will be very large.

If the above explanation is accepted, it means that the changes in frame distances as a function of C_1 are not a reflection of within-class variability but are an artifact of an imperfect alignment process. It means that we have not detected any need to modify the metric as a function of the speech sounds being compared, and it makes it unlikely that the rise in variability at the ends of syllables has anything to do with the kinds of speech sounds predominantly occurring here.

5. BETWEEN-CLASS INFORMATION AND LINEAR DISCRIMINANT ANALYSIS

In addition to the need to scale parameters and resulting distance sensibly, we would like to have a measure of how useful individual parameters and groups of parameters are in the recognition process. For an individual parameter, a way of measuring usefulness is to take the ratio of the scatter in parameter values across different classes to that within the class. When more than one parameter is to be evaluated there is a problem with correlations. A standard method exists for dealing with groups of parameters which involves the simultaneous diagonalization of within- and between-class covariance matrices. The method goes under various names, one of which is linear discriminant analysis [Bricker et al 1971]. Apart from providing a measure of the discriminating power of a set of parameters, linear discriminant analysis can take a set of correlated parameters and linearly transform them into a small set of uncorrelated parameters of equal or similar discriminating power.

We have already described how we derive a within-class covariance matrix. We derive a between-class matrix in a similar way by matching each syllable against a randomly chosen template (actually, the template corresponding to the next syllable in the sentence). In some ways it would be more strictly correct to estimate between-class covariance by substracting within-class covariance from the total covariance, and if what we wanted to do was to estimate the power of a set of parameters to discriminate between all speech sounds equally, then this would be the right approach. However, we ultimately want to improve discrimination among syllables, and syllables do not have speech sounds scattered about them in an even manner: the ends are invariably less loud than the middles and voiceless sounds tend to occur near the edges. We therefore believe that random syllable comparison is a more realistic measure of between-class covariance for our application.

The between- to within-class variance ratios of the individual cepstrum coefficients did not show any obvious pattern, and because of the uncorrelated nature of the original representation, there is little scope for transforming to a representation of lower dimensionality. The linear discriminant analysis approach did not therefore bring any immediate rewards, but it still offers benefits if we wish to include extra parameters which are correlated.

6. TIME-CHANGE INFORMATION AND INTERFRAME CORRELATIONS

The overriding problem remaining in dealing with distance measures is that of spectrum changes with time and interframe correlation: the distance measures coming from adjacent frames in a steady speech sound are not independent and should not be added together as through they were. Moreover, much of the information needed to identify consonants comes from formant trajectories in transition regions. This is particularly true for our low-pass filtered speech, where much of the distinction between the various frication and stop release spectra has been lost. If we knew how to measure real changes in speech sounds and if we knew how to use that information to weight the distance we are summing, there seems little doubt that recognition performance would be greatly improved.

Unfortunately, we do not know how to take time-change information into account. We would first have to know how to distinguish significant differences reflecting changes in vocal tract configuration or manner of production from insignificant differences resulting from the inherent time-variability of voiceless sounds compared with voiced ones, from vocal cord irregularities or from extraneous noises.

As a crude attempt to incorporate time-change information, we tried using the differences between consecutive frames - both the signed differences in individual cepstrum coefficients and the squared Euclidean distance summed over all coefficients - as extra parameters in the recognition algorithm. Part of the rationale was that since the differences between consecutive frames in a steady sound should be close to zero, the difference parameters for any two steady sounds being compared should be close, and the contribution of steady sounds to the total distance between two syllables should be reduced. It turns out, however, that the between- and within-class variances ratios are very small for time difference parameters. It seemed possible that in constructing templates the averaging together of several syllable examples would reduce the random interframe differences and that this might adversely affect the use of difference parameters in recognition. However, average interframe difference in composite templates was not significantly lower than that in the individual examples. We also tried scaling frame differences inversely by syllable length on the grounds that interframe differences would be expected to increase with speaking rate. This,

again, did not produce a positive result. It seems, then, that until a reliable measure of true spectral changes can be found little progress can be made in exploiting spectral change information.

7. SYLLABLE LENGTH INFORMATION

Another kind of information which is not used in the matching algorithm is information about the length of syllables. Table 1 shows that the between-class variance of syllable lengths exceeds the within-class variance by a factor of around 10. Syllable length is therefore clearly a useful source of recognition information (the variance ratios for cepstrum coefficients lie in the range 2 to 4).

	Average Syllable length (Frames)	Between-class Variance (Frames^2)	Within-class Variance (Frames^2)	Variance Ratio
Female French Speaker	38.4	477.5	49.4	9.7
Male English Speaker	32.5	406.3	44.0	9.2

TABLE 1   Syllable Length Data

We would like to incorporate length information with spectral match information in the recognition process. There is, however, a problem in scaling the two sources of information: for uncorrelated variables one can normally just make the within-class variances equal; but spectral information is a property of the frame, and length information of the syllable. Since adjacent frames are highly correlated, we cannot easily know the number of spectral dimensions needed to describe a syllable. Again, we have to wait for progress on the problem of estimating true spectral changes or else determine an empirical scaling factor that optimizes recognition performance. The latter approach would require much more material than is currently available.

Before leaving the topic of syllable lengths, it is interesting to note that, as a proportion of the mean length, the standard deviations of syllable lengths are very close in the two languages. The conventional theory that English is stress timed (approximately equal time intervals between stressed syllables) and French syllable timed (approximately equal time intervals between syllables) would predict a higher consistency in French syllable lengths than in English ones. It is possible that a combination of artifacts in the automatic syllabifier and the restricted syntax limiting the syllable contexts could have obscured the effect, but if it were large it should have shown through.

REFERENCES

Bricker, P.D., R. Gnanadesikan, M.V. Mathews, S. Pruzanski, P.A. Tukey, K.W. Wachter and J.L. Warner. Statistical Techniques for Talker Identification. Bell Syst. Tech. J., Vol. 50, No. 4, pp 1427-1454, April 1971.

Hunt, M.J., M. Lernig and P. Mermelstein. Experiments in Syllable-Based Recognition of Continuous Speech Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Denver, Co., April 1980.

Itakura, F.J. Minimum Prediction Residual Applied to Speech Recognition. IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-23, pp. 67-72, Feb. 1975.

Pols, L.C.W. Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words. Doctoral Thesis, Free Univesity, Amsterdam, 1977.

REPORT DOCUMENTATION PAGE / PAGE DE DOCUMENTATION DE RAPPORT

REPORT/RAPPORT	REPORT/RAPPORT
NAE-AN-57 1a	NRC No. 30144 1b

REPORT SECURITY CLASSIFICATION CLASSIFICATION DE SÉCURITÉ DE RAPPORT	DISTRIBUTION (LIMITATIONS)
Unclassified 2	Unlimited 3

TITLE/SUBTITLE/TITRE/SOUS-TITRE

Distance Measures for Speech Recognition

4

AUTHOR(S)/AUTEUR(S)

M.J. Hunt & C. Lefèbvre

5

SERIES/SÉRIE

Aeronautical Note

6

CORPORATE AUTHOR/PERFORMING AGENCY/AUTEUR D'ENTREPRISE/AGENCE D'EXÉCUTION

National Research Council Canada
National Aeronautical Establishment   Flight Research Laboratory

7

SPONSORING AGENCY/AGENCE DE SUBVENTION

8

DATE	FILE/DOSSIER	LAB. ORDER COMMANDE DU LAB.	PAGES	FIGS/DIAGRAMMES
89/03 9	10	11	37 12a	6 12b

NOTES

13

DESCRIPTORS (KEY WORDS)/MOTS-CLÉS

1. Speech recognition ⌐AN ADA,   3. IMELDA
2. Pattern recognition . ⌐TU⌐,)⌐

14

SUMMARY/SOMMAIRE

This report is concerned with the application of aspects of statistical pattern classification to speech recognition. It presents an extension of linear discriminant analysis to the case where the classes are unknown. This extension provides solutions to the interrelated problems of the design of acoustic representations and spectral distance measures, and allows the efficient combination of heterogeneous sets of parameters. In particular, a representation called IMELDA based on the output of a filter-bank and its changes in time is introduced. Other approaches to distance measures are discussed. It is noted that these other methods lack the ability to make efficient combinations of heterogeneous parameters, and that they require empirical admustments in order to give good results. Tests indicate that IMELDA provides markedly superior recognition performance compared to the alternatives.

15

Page 1473